DOCUMENT RESUME

ED 448 198 TM 032 228

AUTHOR Roberts, J. Kyle
TITLE The Pitfalls of Ignoring Multilevel Design in National Datasets.
PUB DATE 2000-11-16
NOTE 47p.; Paper presented at the Annual Meeting of the Mid-South Educational Research Association (28th, Bowling Green, KY, November 17-19, 2000).
PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Least Squares Statistics; *National Surveys; Regression (Statistics); *Research Design
IDENTIFIERS *Multilevel Analysis; *National Education Longitudinal Study 1988

ABSTRACT

This paper examines the differences between multilevel modeling and weighted ordinary least squares (OLS) regression for analyzing data from the National Educational Longitudinal Study 1988 (NELS:88). The final sample consisted of 718 students in 298 schools. Eighteen variables from the NELS:88 dataset were used, with the dependent variable being the science item response theory estimated number right standardized t-score. Results from the analyses yield no single criterion for choosing one method over the other, but they do illustrate some theoretical situations when multilevel models are preferred. As contextual effects grow larger, multilevel analyses tend to produce more accurate results of the data. Multilevel techniques also allow the researcher to use statistical analyses that are able to mine more complex data. (Contains 2 tables and 29 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

RUNNING HEAD: Multilevel vs. OLS

ED 448 198

The Pitfalls of Ignoring Multilevel Design in National Datasets

J. Kyle Roberts
Baylor College of Medicine

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

J.K. Roberts

Paper Presented at the Annual Meeting of the Mid-South Educational Research Association, November 16, 2000, Bowling Green, KY.

Inquiries concerning this paper should be addressed to the first author at Medical Towers, 1709 Dryden, Suite 545, Houston, TX 77030, (713) 798-7673. Or by e-mail at jroberts@bcm.tmc.edu.

TM032228

BEST COPY AVAILABLE

Abstract

This paper examines the differences between analyzing data from the National Educational Longitudinal Study:88 with two different types of methods; multilevel modeling and weighted ordinary least squares regression. Results from this analysis yield no astounding statistical criterion by which a researcher should choose one method over another, but do illustrate some theoretical situations when multilevel models are preferred.

The Pitfalls of Ignoring Multilevel Design in National Datasets

Introduction

One of the reasons why research has not been overabundant in the investigation of national archived datasets involves the complexity of variables associated with educational issues. Within the last few years, however, the rise of the use of the microcomputer has given software packages the ability to accomplish bigger and more complex tasks with just the click of a mouse. For example, what previously took architects months to draw on blueprints can now be designed in just a day with the aid of AutoCAD. Searches through volumes of literature can now be accomplished in a matter of minutes on the World Wide Web. Statistical procedures that could conceivably take one person years to perform can now be computed in a matter of seconds.

With the aid of computers and complex statistical packages, researchers now have the ability to explore larger and more complex data sets and, in effect, learn more about their fields of study. However, even with the aid of microcomputers, the methods available to researchers may still somewhat limit the scope of questions that they can investigate. For example, even though Spearman conceptualized factor analytic techniques in 1904, it was a technique that was not even attempted until the 1930's and not readily performed until the 1970's with the rise

of the use of the microcomputer and complex statistical packages. Now, researchers with the aid of SPSS, SAS, and other statistical software use factor analysis routinely.

The same limitations have emerged with the development of multilevel design, a promising approach for the study of the complex relationships we encounter in education. Almost half a century ago, Robinson (1950) discovered the need for multilevel techniques while performing regression analyses at different levels of variables (i.e., regressions with students and regressions with schools). Later termed the *Robinson effect*, these different level regressions "[showed] that analyses executed at different levels of the hierarchy do not necessarily produce the same results" (Kreft & de Leeuw, 1998, p. 3). Although these regressions often gave opposite results when measured at different levels, no statistical method existed that could overcome this problem.

The problem that Robinson was facing was the ability to describe data that have group regressions with both random slopes (differences among schools) and random intercepts (differences among students). This problem occurs in many large-scale data sets (Seltzer, 1994). The challenge is two-fold; it is necessary to not only recognize the need for multilevel techniques, but to also utilize the potential value

of multilevel techniques in broadening the types of questions that can be addressed (Seltzer, 1991).

Despite the apparent promise of multilevel design, few researchers have used these techniques to study complex problems such as science performance in urban schools. Just as Robinson (1950) first noted, companions of more commonly used methodologies invoking multilevel methods might reveal differences in findings across units of analysis that have implications for both policy and practice.

In the present field of study, there is a distinct need for multilevel techniques. Because the focus of the present study is to examine differences between results obtained from multilevel analyses and ordinary least squares (OLS) analyses, it seems reasonable that any data that might be analyzed within the study should be treated as level-1 variables (students) nested within level-2 variables (urban schools). Treating the data this way will allow researchers to both identify schools that have students performing at different levels (mean/intercept differences) and have greater rates of learning (slope differences).

The use of multilevel techniques, instead of OLS methods, will help in the interpretation of results. Instead of fitting just one regression line to the data, multilevel techniques recognize that the data are nested into groups and give

researchers an understanding of where and how effects are occurring (Goldstein, Rasbash, Plewis, Draper, Browne, Yang, Woodhouse, & Healy, 1998). Using multilevel techniques allows the researcher to estimate the pattern of variation of the schools. This is of great benefit because it helps to identify variables that have a great amount of variation; that is, variables in which there are large differences between students within schools. Ignoring this clustering effect may cause the standard errors of the regression (OLS) to be underestimated. Specific to the field of science education in urban schools, multilevel techniques provide greater ability to identify variables in which urban schools differ greatly (complex level-2 variation) when regressed on the student science achievement scores.

The present study will attempt to examine differences between these two types of analyses (OLS and multilevel) by investigating an archived dataset, the National Educational Longitudinal Study of 1988.

NELS:88 Overview

While techniques used for affecting student achievement have flourished under recent research, methods for uncovering the identifying characteristics of successful students in successful schools has also improved. One of the main reasons

for this improved identification of predictor variables is the improvement in the quality of data collection techniques. One such result of successful data collection is the National Educational Longitudinal Study (NELS:88). What follows is a brief description of the content of the NELS:88 dataset, an overview of the instrument development, validity and reliability reports, and methods for dealing with the complexity of the data.

Background of the NELS:88

The NELS:88 project was headed by the National Center for Educational Statistics of the U.S. Department of Education. The study was begun in February of 1986 as part of an extension of the information gathered from other Department of Education tests such as the National Longitudinal Study (NLS-72) and High School and Beyond (HS&B). Unlike NLS-72 and HS&B which focused on high school seniors and their post-secondary education, the NELS:88 was designed to follow eighth graders from 1988 through 1994, measuring their academic progress at two year intervals.

The NELS:88 was originally designed to produce a general purpose data set that could be used to inform policy makers on current trends and needs for reform (Ingels, Scott, Lindmark, Frankel, & Myers, 1992). Some of the policy issues that the NELS:88 attempted to answer were the "identification of school attributes associated with achievement, the transition of

different types of students from eighth grade to secondary school, the influence of ability grouping on future educational experiences and achievements, determinants of dropping out of the educational system, and changes in educational practices over time" (Ingels et al. 1992, pp. 5-6). Through the use of an extensive parent questionnaire, the NELS:88 also provided insights into the role of parent(s) in the student's education, something that other national data sets had not included.

Large national databases, such as the NELS:88, are able to provide researchers with an added strength over independently commissioned studies because the databases incorporate a larger and more accurate population sample. As a result, these national data sets produce comprehensive data that are more effective in gauging the effectiveness of existing school-based programs and reform efforts. Due to the longitudinal design of these data sets, researchers are also able to analyze trend data, which can aid in pointing to the most critical experiences of high school students (Haggerty et al., 1996).

Undertakings such as the NELS:88 are simply not feasible for a single researcher or even team of researchers. National data set collections often have several commodities that independent researchers do not possess: money, time, and access. Because of constraints of budget, independent researchers are rarely able to collect a sample population as

large, as long (duration), or as geographically widespread as the NELS:88. As a result, smaller research initiatives tend to produce samples that are less representative of the population.

Although national data sets allow many research opportunities, they also have disadvantages. First, and most importantly, is the mismatch between researcher intent and the instrument's contents. Many of the questions that are addressed by external researchers utilizing these datasets are not questions that were of primary interest to the developers of the instrument(s). The difficulty with using the NELS:88 and other national data sets is that researchers are trying to answer their own questions (primary investigation) with someone else's data (secondary analysis).

Instrument Development

The base year of the NELS:88 instrument covered content categories including constitutional factors (sex and age), ethnicity, home characteristics, socioeconomic status, work status, attitudes and values, school characteristics, school atmosphere, school work, school performance, guidance, special programs and after-school programs, involvement with community, life goals, and financial assistance (Ingels et al., 1992).

The NELS:88 was not just concerned with collecting data only from the student's perspective. In an effort to provide contextual sources for student outcomes, parents, teachers, and

administrators were also surveyed. The school administrator questionnaire was designed to collect information on school characteristics, policies and practices, grading and testing structure, parent involvement, and school climate (Ingels et al., 1992).

The primary purpose of the teacher questionnaire was to provide teacher information that could be used to analyze both behaviors and outcomes of the student sample. These surveys were administered to two of each sample student's teachers in two of the four cognitive areas covered by the student questionnaire (i.e., mathematics, science, reading, and social studies).

The NELS:88 instrument was developed not only to provide current insight into the state of education, but also to allow for cross-cohort research with other longitudinal data sets such as HS&B and NLS-72. This aspect of the NELS:88 allows researchers to conduct trend analysis between high school sophomores (NELS:88 and HS&B) and high school seniors (NELS:88, HS&B, and NLS-72).

Sampling Design and Issues

The NELS:88 sampled approximately 1,000 schools from the over 40,000 public and private schools in the United States. Within each of these schools, 24 eighth-grade students were randomly selected to represent the nearly 3,000,000 students in

schools in 1988. Among these 24 students for each school, an additional 2-3 Asian and Hispanic students were added in over-sampling to allow for generalization concerning policy relevant groups.

Types of Data/Questions Available

Although the NELS:88 was designed to investigate a wide range of research questions, researchers must ensure that their questions using the NELS:88 are adequately represented in the data. Some of the research issues that can be addressed by the NELS:88 include, but are not limited to:

1. Students' academic growth over time.
2. Transition from eighth grade to high school.
3. The process of dropping out of school, as it occurs from eighth grade on.
4. The role of the school in helping the disadvantaged.
5. The school experiences and academic performance of minority students.
6. Students' pursuit of the study of mathematics and science.
7. The features of effective schools.
8. Access to and choice of postsecondary schools.
9. Transitions to postsecondary education and the world of work.
10. Trend analyses with previous longitudinal studies (NCES, 1999a).

Psychometric Properties/Issues of the NELS:88: Validity and Reliability

As with any type of self-report survey, issues of score validity and reliability are always a concern. In an article in *American Educational Research Journal*, Nussbaum, Hamilton, and Snow (1997) examined issues related to the validity of assessment scores like the NELS:88, specifically in relation to science assessment. They stated that part of the difficulty of interpreting results from these broad surveys is that data are often "limited to fairly superficial description" (p. 168). Among their findings on the NELS:88, they discovered the following:

1. The eighth-grade science achievement scores seem to be ambiguous and unstable.
2. This instability in the eighth grade seems to consolidate and stabilize as students progress through high school.
3. The NELS:88 science tests are actually multidimensional, measuring three factors: quantitative science; spatial-mechanical reasoning; and basic knowledge and reasoning. Analyses based on total scores often misses important effects.
4. Part of the reason for unreliable scores from eighth-grade students can be explained by the fact that "middle school science courses are general and heterogeneous, with diverse

and nonstandardized content, relative to the more specialized high school courses" (p. 169).

These results were consistent with the findings of Rock, Pollack, and Quinn (1995) who conducted tests of the reliability of the IRT Theta "T" score. Overall, they discovered that the theta for the science measures was consistently lower than those of the math and reading measures. They also discovered that the reliability of theta for the base year science measure (.73) was lower than both the first and second-year follow-ups (.81 and .82, respectively).

With respect to the NELS:88, it should be noted, that results from the achievement tests tend to produce more reliable data as the students move further through their educational experience. Therefore, if analyses are to be conducted using achievement scores as outcome variables, a researcher would do better to use the tenth- and twelfth-grade scores rather than the eighth-grade scores.

Although it may seem that, when looking at science achievement in the NELS:88, a researcher should use the eighth grade science scores as a last resort, several objections must be raised to the argument posed by Nussbaum et al. (1997). The first objection is the issue of sample size. With complicated data analysis techniques, larger sample sizes often afford a more accurate interpretation of interactions within the data.

This is especially true with multilevel modeling. When a researcher decides to use either the tenth- or twelfth-grade sample from the NELS:88, the student within-school count drops. Part of the reason for this is that students have either moved out of a district or have been funneled into a different high school than their peers. For example, in the base year of the NELS:88, one school may have had 10 students who were sampled out of that school class of 100 students. Although it would be expected that all of these students would attend the same high school, two may have moved in the two years before the first follow-up was collected. The student mobility then effectively cuts the within school sample from 1:10 (10:100) to 2:25 (8:100). And because most urban schools have a high mobility rate among their students, researchers must consider the research design and ask whether or not the increased alpha in science scores is worth the decreased student within-school count.

A second issue to consider is the structure of science courses in high school versus in eighth grade. Nussbaum et al. (1997) noted "the relationship between course taking and ability is probably reciprocal" (p. 169). This means that although students who take more science courses will have higher science achievement, the composition of students within those classes are often students who already excel at science, and

consequentially enjoy taking more science courses. As a result, at the high school level, student achievement in science is probably more of a byproduct of initial student ability rather than school-level factors. The eighth-grade sample can provide some help in overcoming this problem because middle-school science curriculum is often more homogeneous across schools, especially within states.

Third, the objections posed by showing that the science tests actually measure three different constructs do not apply as readily to the eighth-grade sample. Nussbaum et al. (1997) noted "a reliable factor structure [arguing for three independent factors] emerges in tenth grade and twelfth grade" (p. 171). The factor structure is not nearly as strong for the eighth-grade sample. Therefore, if a researcher were to use a single factor from the science IRT theta T-scores, an argument could be made that the eighth grade sample best represents a test measuring the single construct of science achievement.

Weights, SE, and Design Effects

In an effort to compensate for the student nonresponse and unequal sampling probabilities, NELS:88 has a series of weights built into the data set. Haggerty et al. (1996) described the weighting process as involving two stages.

> In the first step, unadjusted weights are calculated as the inverse of the probabilities of

> selection, taking into account all stages of the sample selection process. In the second step, these initial weights are adjusted to compensate for unit nonresponse; such nonresponse adjustments are typically carried out separately within multiple weighting cells. (p. 5-1)

Failure to use weighted samples with the NELS:88 could result in two fallacies: under-representation and over-representation. As was mentioned earlier, the NELS:88 over-sampled some students (Hispanics, Asians, and private school students) in an effort to provide better data for some sub-population analyses. Failure to use weights with these students will result in over-representation in the data. At the same time, other students, because of nonresponse and under-sampling, are under-represented in the data. For example, because of under-representation in the second follow-up, some students have a weight of 6,670 (NCES, 1999). This means that, because of either nonresponse or under-sampling, that one student represents 6,670 other students versus an original weight of 120.

When the NELS:88 was collected, the data were not just a random sample from a population of students. Instead, the sample design involved the disproportionate sampling of certain groups/strata and clustered (multi-stage) probability sampling.

The consequence of this data collection method is that the resulting statistics are more variable than if they had been drawn from a random sample of the population. As a result, analyses cannot simply be performed on the data set assuming that the variability of this data would be the same as that in a random sample. Because of the data collection design, the data in the NELS:88 is more variable than data collected from a simple random sample (Haggerty et al., 1996). Therefore, correct standard errors must be computed before analyses are run. Procedures for calculating correct standard errors include Taylor Series Approximations, Balanced Repeated Replication, and Jackknife Repeated Replication.

Missing Data

One thing not addressed by the weighting of variables or the correction of standard errors and design effects is the issue of student drop off and student mortality in the subsequent follow-up administrations of the NELS:88. In order to address this issue, the three follow-up surveys included "freshened" students, "base-year ineligible" students, and subsampling. The freshened students were additional tenth graders and seniors who were not part of the original sample, but were added so that in subsequent years follow-ups would be representative samples.

The base-year ineligible students were individuals who were deleted from the original sample (1988) by the school principal for reasons of disability. In the first and second follow-ups, these students were added back into the sample if it was felt that their condition no longer presented a hindrance to data collection or to the sample.

Another difficulty that the developers of the NELS:88 had to overcome in the first and second follow-up was tracking the almost 25,000 eighth grade students in 1000 middle schools to almost 5000 high schools. Because some of these high schools enrolled few NELS:88 students, a decision was made to subsample these students. Students included in the subsample fall into one of two categories: (1) students who transferred out of their original school; and (2) nonrespondents who were originally classified as potential dropouts. From the transfer and "potential dropout" students, a 20% subsample and 50% subsample, respectively, was drawn in the first follow-up.

However, when using multilevel techniques over non-multilevel (OLS) techniques, missing data becomes less of an issue. Multilevel analyses do not require nor assume that data are completely crossed/balanced. Instead, the estimation procedures are based on the assumption that "the probability of being missing is independent of any of the random variables in

the model" (Goldstein et al., 1998, p. 61). Goldstein et al. (1998) go on to explain this procedure:

> [All available] data can be incorporated into the analysis. . . [The condition of being missing], known as completely random dropout may be relaxed to that of random dropout where the missing mechanism depends on the observed measurements. In this latter case, so long as a full information estimation procedure is used, such as that of maximum likelihood . . ., then the actual missingness mechanism can be ignored. (p. 61)

For the researcher, this means that when repeated measures analyses are preformed in multilevel modeling, there is no need to drop students who do not have multiple measurements. Further discussion of multilevel technique's ability to handle missing data is presented by Diggle and Kenward (1994).

Multilevel Modeling

The history of multilevel modeling can be linked back to the seminal work of Robinson (1950) in recognizing contextual effects. Robinson's discovery was not unlike the differences noted while neglecting structure (or context) in ANOVA designs when using a crossed/balanced design over a nested design (Roberts, 2000). Neglecting the fact that individuals or

measurement occasions may be nested inside other larger clusters will often lead researchers to erroneous conclusions about their data.

For illustrative purposes, let us suppose that a researcher is interested in English proficiency among students within schools across Texas. Simply performing an ANOVA to test for differences between mean English proficiency scores within the schools would neglect the fact that some of the schools closer to the border of Mexico have a larger number of non-English speaking students. Neglecting this structure might lead a researcher to assume that a school is doing a poor job of educating its students in English, when in fact they are doing a superb job of teaching English, with respect to the population of students in their school.

If percentage of non-English speaking students was then considered as a covariate and an ANCOVA design was used, a researcher would then be able to discover whether or not schools differed. Although the ANCOVA method begins to correct some of the problems associated with neglecting group structure, it still can only provide answers to the question of if schools differ and not why schools differ (Kreft & de Leeuw, 1998). In a sense, multilevel analysis combines the strengths of regression and ANCOVA designs by allowing researchers to predict outcome scores with other continuous or non-continuous variables

while taking into account the fact that the scores may be nested within groups. Hence, we are able to not only determine which schools differ, but examine why they differ.

What are Multilevel Models?

Although the popularity and awareness of multilevel and hierarchical linear models has increased dramatically in the last few years, it would be helpful here to provide a definition and primer of these techniques. Multilevel statistical models (MLM) may be regarded as an extension of the General Linear Model (GLM). The GLM subsumes most statistical techniques like ANOVA, ANCOVA, MANOVA, regression, and canonical correlation (Fan, 1996). The advantage of multilevel modeling over simple regression or ANOVA is that it allows the researcher to look at hierarchically structured data and interpret results without ignoring these structures. This is accomplished in MLM by including a complex random part which can appropriately account for correlations among the data.

Among the statistical packages currently developed for running multilevel procedures, three are most commonly used: (1) MLwiN, developed by Harvey Goldstein and the staff at the Multilevel Models Project, Institute of Education, University of London (Goldstein, Rasbash, Plewis, Draper, Browne, Yang, Woodhouse, & Healy, 1998); (2) HLM, developed by Bryk, Raudenbush, and Congdon (1996); and (3) PROC MIXED, a routine of

the SAS statistical package (Singer, 1999). For the purposes of this paper, multilevel analyses will be illustrated by MLwiN. While each package has differing strengths, MLwiN was chosen because of its notation and graphing capabilities. For a more detailed discussion of the software packages available for multilevel analysis, see Kreft and de Leeuw (1998) and Kreft, de Leeuw, and van der Leeden (1994).

A multilevel or hierarchically structured dataset can take many forms. All that is required is that level-1 units of some type (e.g., students or measurement occasions) be nested inside level-2 units (e.g., schools or years). Although the two-level structure is the most common, multilevel models are not restricted to just two levels, they simply must have at least two levels.

Consider the following examples of multilevel data sets: students nested within classrooms, students nested within schools, students nested within classrooms nested within schools, people nested within districts, measurement occasions nested within subjects (repeated measures), students cross-classified by school and neighborhood, and students having multiple membership within schools across time (longitudinal data). Each of these examples illustrate data that are considered hierarchical in structure. Data derived from such

hierarchical designs may be correlated, and the analysis must take this into account.

Failure to recognize hierarchical data structures and implement multilevel techniques could also result in misinterpretation in the analysis. First, statistical models that are not hierarchical sometimes ignore the structure of the data and as a result report underestimated standard errors (no between-unit variation), thus resulting in increased Type I errors. Goldstein et al. (1998) illustrate that when using OLS methods, one course of action may be chosen over others when in fact that course may be due solely to chance.

Second, multilevel techniques are much more statistically efficient than other techniques. Suppose that a researcher wanted to explore the difference between math scores and SES among 10,000 students in 300 schools. In order to look at the different school effects, the researcher would be forced to plot 300 regressions (one for each school) and then attempt to interpret results based on these regressions. Multilevel techniques are more efficient because they do not require the researcher to estimate all of these effects. Goldstein et al. (1998) also note that "because [the OLS equation] does not treat schools as a random sample it provides no useful quantification of the variation among schools in the population more generally" (p. 12).

Third, multilevel techniques assume a general linear model, and as such, can perform multiple types of analyses that provide more conservative estimates by allowing for correlated responses within clusters.

As was alluded to earlier, multilevel modeling can perform an array of analyses including ANCOVA, regression (OLS and GLS), maximum likelihood estimation, repeated measures, meta-analysis, multivariate response, Bayesian modeling, binary response, bootstrap estimation, and multiple membership models. Although all of these methods are available options to the researcher using multilevel modeling, this paper will primarily deal with the random effects model that can be used to analyze data obtained from students nested within schools.

Intraclass correlation

Intraclass correlation (ICC) is the proportion of total variance that is between the groups of the regression equation. Put more succinctly, it "is the degree to which individuals share common experiences due to closeness in space and/or time" (Kreft & de Leeuw, 1998, p. 9). Hox (1995) explains the ICC as a "population estimate of the variance explained by the grouping structure" (p. 14). This concept is important to the researcher because if intraclass correlation exists, then the traditional linear model must be abandoned because the assumption of

independent observations has been violated (Kreft & de Leeuw, 1998).

In a two-level model, the ICC is found by dividing the variance at the highest level (in this case the level-2) by the sum of the variances at the lowest level and the highest level. In other words, as Equation 1 explains, ICC (ρ) for a two-level model is the proportion of group level variance from the total variance, where σ^2_{u0} represents the level-2 variance and σ^2_{e0} represents the level-1 variance.

$$\rho = \frac{\sigma^2_{u0}}{\sigma^2_{u0} + \sigma^2_{e0}} \qquad (1)$$

It is helpful here to illustrate the importance of ICC. Let us suppose that a researcher has collected data on science achievement from four schools where one is urban, one is suburban, one is private, and one is rural. The traditional OLS model would assume that each of these observations was independent of the context/school in which the data were collected, therefore neglecting intraclass correlation. Thus, the prediction of student scores in science achievement would be estimated irrespective of the type of school that the student attended.

The multilevel model allows the possibility that the students' scores on a given outcome variable may be partly a function of the school that they are in. Students within the

same school may tend to be more alike than students in different schools, thus causing a greater dependency of observations, or high intraclass correlation. Thus the presence of a high intraclass correlation would mean that the highest level of the predictor variable should be modeled as random to reflect the fact that students tend to be more like students in their own school rather than students in other schools.

Power of the ICC and the Design Effect

In the school effects model, intraclass correlation is often the first statistic consulted when determining the amount of total variance attributable to the differences between schools. However, it is also important to consider the design effect. For a two-level model, the design effect is computed with the following formula:

$$Deff = 1 + (B-1)*\rho \qquad (2)$$

where ρ is an estimate of the intraclass correlation and B is the cluster size (or average cluster size) (Snijders & Bosker, 1999).

Once the design effect has been computed, it can be used to approximate the effective sample size given the actual sample size. This is somewhat similar to performing a "what if" analysis in regression and ANOVA. The purpose in computing the design effect is to determine the statistical power of the design given the actual sample size, cluster size and ICC. Once

the design effect has been computed, the effective sample size (Ne) can be computed with the following equation:

$$Ne = N/Deff \quad (3)$$

Therefore, the statistical power of the ICC would depend on the cluster size. For example, if either B = 1 or the ICC = 0, then Deff = 1 and Ne = N. In this case, the ICC has low statistical power.

Therefore, a large intraclass correlation would be one that, given the cluster sizes, would reduce the effective sample size (Ne) below some acceptable threshold. Although not an estimate of power, the design effect of an ICC is helpful in determining whether or not the researcher needs to model random level-2 variables in the regression equation. Snijders and Bosker (1999) report that most educational research reports ICC values between 0.05 and 0.20. Values greater than 0.20 for the ICC should be considered large. For a discussion of the relativeness of power and sampling sizes, see also Snijders & Bosker (1993) and the discussion of the PINT (Power IN Two-level designs) statistical package.

Statistical Significance

Although statistical significance is often one of the first things determined and reported in univariate and multivariate analyses, it has come under a considerable amount of fire recently (Thompson, 1998). While obtaining a p-value below 0.05

is often considered the mark of "significant" findings in the linear equation, multilevel modeling is rarely concerned with obtaining p-value estimates, but instead concerned with the power of the multilevel analysis.

Using statistical significance techniques is sometimes helpful, however, when determining the relative strength of the influence of predictors. When trying to decide whether or not to free or fix parameters for a multilevel model, sometimes chi-square-versus-degrees-of-freedom tests are used to examine the difference between two models. Using chi-square-versus-degrees-of-freedom tests not only helps determine which models differ significantly, but also helps the researcher produce a parsimonious model.

For example, suppose a researcher decides to free an explanatory variable at level-2 and in doing so adds four degrees of freedom to the model. After this new model is run, the researcher discovers that the change in chi-square (or -2*log likelihood) is only 2.1. Thus, the chi-square-versus-degrees-of-freedom for 2.1 on 4 degrees of freedom is not statistically significant. These results indicate that the given variable probably should not be freed at the school-level (level-2).

Although this type of testing is frequently used in multilevel modeling to determine model fit and make decisions

about parsimony, it requires that the model be nested. For a more detailed discussion of other tests that may be performed to test for differences between models, see Bryk and Raudenbush (1992, pp. 48-59).

Variance Explained

Often times in the OLS regression model and in other OVA models, researchers are concerned not only with whether or not their predictors are statistically significant, but also the amount of variance explained (R^2) by each predictor and by the total model. Once again, however, this is often not the case in multilevel modeling. As was mentioned previously, the purpose of multilevel modeling is on estimating the pattern of responses across schools.

Determining the amount of variance explained in a multilevel model is a very difficult process. For the multilevel model, the variance explained is divided into the variance accounted for at each level of the hierarchy. When computing the variance explained for the two-level model, the level-1 R^2 can be found by dividing the variance of the empty model (a random effects ANOVA model with only the general mean, random groups, and random variations within groups) by the variance for the full model (all predictors included) and then subtracting that from one.

$$1 - \frac{\sigma^2_{e0} + \sigma^2_{u0}empty}{\sigma^2_{e0} + \sigma^2_{u0}full} \quad (4)$$

Level-2 variance explained is then computed by dividing the level-1 variance by the cluster size, as is illustrated in Equation 5.

$$1 - \frac{\frac{\sigma^2_{e0}}{B} + \sigma^2_{u0}empty}{\frac{\sigma^2_{e0}}{B} + \sigma^2_{u0}full} \quad (5)$$

Researchers should be cautioned from interpreting the amount of variance explained, however, because adding predictors can sometimes lead to a negative R^2 if the variable added increases the amount of variance at one level or another (Snijders & Bosker, 1999).

While Snijders and Bosker (1999) partition out the variance explained at each level, the variance explained by the overall multilevel model can also be computed as follows. The first step in this procedure is fitting the multilevel model in the usual way. The predictions, or $\tilde{y}$, are calculated using the empty model, e.g., using only the grand mean as predictor. The predictions $\hat{y}$ are then calculated using the full model, e.g., adding all predictors. The sum of squares total is then $(y - \tilde{y})^2$, and the sum of squares error is $(y - \hat{y})^2$. The sum of squares explained can then be calculated by subtracting the SST from the SSE. Finally, the R^2 is calculated by (SSR/SST) * 100%.

Although it is helpful to compute R^2 for the purposes of comparing OLS models in terms of variance explained, an R^2 will not be computed here since the underlying comparison is between the OLS and multilevel models.

Analysis

Sample

As was mentioned previously, the sample was drawn from the base year of the NELS:88, which contained 24599 students. Of these students, 7620 were in urban schools, 10246 were in suburban schools, and 6733 were in rural schools. Because the present study was only concerned with differences between performance of urban schools, the suburban and rural school students were dropped from the analysis. Of these 7620 students, 307 did not have student tests available, thus reducing the sample to 7313 students in 317 schools. Following the guidelines set forth by Lawrence and McLean (1999), the dataset was further reduced to include only schools that had at least 10 students within the school. This was done in an effort to maintain robustness of estimation in multilevel modeling.

The final sample consisted of 7178 students in 298 schools. Of these students, 597 were Asian/Pacific Islanders, 1339 were of Hispanic origin, 1432 were African American, 3624 were Caucasian, and 84 were American Indian or Alaskan Native. When

applying the sampling weights to these observations, a weighted sample of 704786 was used for analyses.

Variables Extracted

For the purposes of the present study 18 variables were extracted from the NELS:88 dataset and examined. Each of these variables were chosen based on current research (Boyd & Shouse, 1997; Hoffer et al., 1996) and the applicability of these variables in describing the differences between urban school students' achievement on science outcome variables.

The outcome, or dependent, variable selected for the present analysis was the science item response theory (IRT; Fan, 1998) estimated number right standardized t-score (BY2XSSTD). Although the NELS:88 provides many science outcome variables to examine (science IRT estimated number right, science standardized score, science percentile, and overall science proficiency), the IRT estimated scores were chosen for two reasons. First, unlike some of the science proficiency estimates, the IRT estimated score is a continuous variable and adds variance which maintains outcome variable in the analyses.

Second, when dealing with students at extremes of the distribution, IRT estimates are traditionally better predictions of student success than standardized or raw scores (Fan, 1998; Lawson, 1991). Instead of simply reporting the number correct, or raw score, of a student on the science achievement test, the

IRT estimates instead considers the pattern of scores for that individual and assign a score based on the pattern of responses.

One problem that researchers face when choosing to use the IRT estimated scores, however, is just which IRT estimate to use. The NELS:88 reports three different IRT scores: IRT estimated number right raw score metric, IRT estimated number right standardized t-score, and the IRT theta t-score. When examining cohorts at one point in time, Ingels, Dowd, Baldridge, Stipe, Bartot, and Frankel (1994) recommended using the IRT estimated number right standardized t-score, because it has been standardized within years, as opposed to the IRT theta t-score, which is standardized between years.

The necessary weight for this selection of students was the base year student questionnaire, BYQWT. Because this weight was being used, traditional software packages such as SPSS and SAS had to be abandoned. For purposes of this paper, the SUDAAN software package (Shah, Barnwell, and Bieler, 1997) was selected. Using SUDAAN will also require the use of the Superstratum ID variable (SSTRATID) so that SUDAAN could correctly estimate the standard errors for the dataset. As was previously mentioned, NELS:88 sampling techniques oversampled certain subgroups of students. As a result, one Hispanic student may represent only 600 students, while one Caucasian student may represent over 2000 students.

SUDAAN allows for the correct computation of standard errors and variance estimation by using both replication methods and Taylor series linearization for obtaining variance estimates of both the descriptive statistics and regression parameters (Shah, Barnwell, & Bieler, 1997). There is a slight problem in the apples-to-apples comparison of the SUDAAN and multilevel modeling procedures in that the dataset used for the multilevel model will not be weighted. Although the method of using weighted samples when computing estimates in MLwiN or HLM, these software packages do not currently allow for the inclusion of weights.

Results

Table 1 presents the results from the single predictor regressions for both the multilevel and weighted samples. For each of the analyses, β_0 represents the intercept and β_1 represents the unstandardized slope. The ICC, or intraclass correlation, is also presented in this table.

Insert Table 1 about here

Results from Table 1 were then analyzed in a linear regression where the gain scores in slopes (absolute value) was defined as a dependent variable and the ICC was defined as the

independent variable. The results of this linear regression yielded non-significant results ($p = 0.760$) with an R^2 of 0.006 ($F = 0.096$).

As can be seen from Table 1, there seems to be no discernable pattern for determining which absolute value of the slope will be greater. In this sample, 11 of the slopes for the OLS sample were greater (absolute value) and 7 of the slopes for the multilevel sample were greater (absolute value).

Several of these variables were also included in a multiple regression in both a weighted OLS and multilevel model to determine if there are any discernable patterns when multiple predictors are included in an equation. The results are presented in Table 2.

Insert Table 2 about here

However, as was noted from the results of Table 1, the results of the multiple regressions yield no discernable patterns either. Discussions of the implications of these findings will be presented in the following section.

Discussion

Although no statistical considerations for the implications of not using multilevel analyses with the NELS:88 seemed to come

to the surface, some practical considerations should be discussed here.

Results from Tables 1 and 2 seem to provide some means of insight into interpreting differences between OLS and multilevel results. As was mentioned previously, OLS models sometimes ignore the structure of the data and as a result report underestimated standard errors (no between-unit variation), thus resulting in increased Type I errors. Although the underestimated standard errors will potentially affect the estimation of the weights for the slope coefficients, there can be no independent way of determining how much (or in what direction) the slope will be affected.

By looking at the results of Table 2 with some qualitative assessment, a small pattern seems to begin developing. The two variables that have the greatest magnitude of difference in slope coefficients (BYSC45B3 and WHITE) are both variables that would seem to have strong contextual effects. In this case, the variable BYSC45B3 (science taught in a non-English language) is scored "1" = yes and "2" = no. The results from this multiple regression would seem to show that this variable is measuring an artifact of the number of students in the school who are non-English speakers and probably recent immigrants to the US. The OLS estimate, which would present an "average" slope coefficient across all schools, shows that there is only a slight difference

between these schools. The multilevel estimate instead shows that given the context of the school, the difference between students within schools is actually much larger. When further investigation of this variable was carried out, it was also discovered that the standard error for β_1 for the multilevel estimate was twice as large as the estimate for the OLS equation (0.86 and 0.42 respectively).

The variable WHITE (White students versus the rest of the students) would also seem to be a variable that would have strong contextual effects. This would seem even more contextual given the fact that this sample includes only the urban schools. What this strong difference in results could indicate is that many white students attend higher achieving schools which are simply placed in urban settings (e.g., private schools, urban high income schools, etc.).

Given these qualitative inquiries, the only conclusion that can be immediately drawn is that the larger the contextual effects, the more multilevel models are needed over OLS models. These findings are corroborated by Kreft and de Leeuw (1997), Roberts (in press), and Snijders and Bosker (1999).

Conclusion

This is somewhat a frustrating paper to both write and read (I imagine). It would seem that there are no real spikes of

truth that could be concluded from this paper. While this could be argued, it should be pointed out that the general purpose of this paper was to compare and investigate the differences between a weighted OLS strategy for analyzing the NELS:88 and a multilevel strategy. Results do show differences, but no discernable pattern about these differences.

The strength that I see in using multilevel strategies over OLS are threefold. First, as contextual effects grow larger, the multilevel analyses tend to produce more accurate results of the data. This was illustrated with the data presented in the multiple regressions in Table 2.

Second, multilevel techniques allow the researcher more statistically savvy analyses which are able to mine more complex data. An example of this would be the analyzing of complex cross-classified data and trend data where students have multiple membership in different schools. Being that some of the questions that can be asked across years with the NELS:88 require the use of complex techniques, multilevel methods seem preferable when working with this dataset.

Finally, multilevel techniques (and specifically the MLwiN software package) will allow for the identification of high achieving schools when the focus of the study is a continuous outcome variable. Some of the extended graphing capabilities allow researchers to plot residuals and then identify schools

with greater rates of increase in learning than other schools. This can be especially helpful when trying to identify models for school reform. Other packages must deal with this as either an ANOVA (non-continuous outcome variables) or as a single least squares line prediction.

Although there has been a slight case made here for the utility and use of multilevel modeling over OLS, it should be pointed out that these procedures are difficult and have a steep learning curve. Researchers should be cautioned from simply applying multilevel techniques. Some datasets often call for more complicated methods such as modeling variance and error at different levels of the hierarchy. In cases where researchers are unsure of the application of multilevel methods, OLS techniques should be utilized.

References

Boyd, W. L., & Shouse, R. C. (1997). The problems and promise of urban schools. In H. J. Walberg, O. Reyes, & R. Weissberg (Eds.), *Children and youth: Interdisciplinary perspectives* (pp. 141-165). Thousand Oaks, CA: Sage.

Bryk, A. S., Raudenbush, S. W., & Cogdon, R. T. (1996). *HLM. Hierarchical linear and nonlinear modeling with the HLM/2L and HLM/3L programs*. Chicago, IL: Scientific Software International.

Bryk, A. S. & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods*. Newbury Park, CA: Sage.

Diggle, P. & Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *J. Royal Statistical Society, C*(43), 49-93.

Fan, X. (1998). Item response theory and classical test theory: An empirical comparison of their item/person statistics. *Educational and Psychological Measurement, 58*, 357-381.

Fan, X. (1996). Canonical correlation analysis as a general analytic model. In B. Thompson (Ed.), *Advances in social science methodology, vol. 4* (pp. 71-94). Greenwich, Connecticut: JAI Press.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G., & Healy, M. (1998). *A user's guide to MlwiN*. Multilevel Models Project, Institute of Education, University of London.

Haggerty, C., Dugoni, B., Reed, L., Cederlund, A., & Taylor, J. (1996). *National education longitudinal study: 1988-1994* (NCES 96-174). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Hoffer, T. B., Moore, W., Quinn, P., & Suter, L. E. (1996). *High school seniors' instructional experiences in science and math* (NCES 95-278). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Hox, J. J. (1995). *Applied multilevel analysis*. Amsterdam: TT-Publicaties.

Ingels, S. J., Dowd, K. L., Baldridge, J. D., Stipe, J. L., Bartot, V. H., Frankel, M. R. (1994). *User's manual: National educational longitudinal study of 1988 second follow-up: Student component data file user's manual* (NCES 94-374). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Ingels, S. J., Scott, L. A., Lindmark, J. T., Frankel, M. R. & Myers, S. L. (1992). *National education longitudinal study of 1988 first follow-up: Student component data file user's*

manual (NCES 92-030). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Kreft, I. & de Leeuw, J. (1998). *Introducing multilevel modeling.* Thousand Oaks, CA: Sage.

Kreft, I., de Leeuw, J., & van der Leeden, R. (1994). Review of five multilevel analysis programs: BMDP-5V, GENMOD, HLM, ML3, VARCL. *American Statistician, 48,* 324-335.

Lawson, S. (1991). One parameter latent trait measurement: Do the results justify the effort? In B. Thompson (Ed.), *Advances in educational research: Substantive findings, methodological developments* (Vol. 1, pp. 159-168). Greenwich, CT: JAI Press.

National Center for Education Statistics (1999, September 17). *Responses to frequently asked questions regarding NELS:88* [WWW document]. URL http://nces.ed.gov/surveys/nels88/ques.shtml

Nussbaum, E. M., Hamilton, L. S., & Snow, R. E. (Spring, 1997). Enhancing the validity and usefulness of large-scale educational assessments: IV. NELS:88 science achievement to 12th grade. *American Educational Research Journal, 34,* 151-173.

Roberts, J. K. (in press). An Introductory Primer on Multilevel and Hierarchical Linear Modeling. In B. Thompson (Ed.), *Advances in Social Science Methodology (Vol. 6).* Stanford, Conn: JAI Press.

Roberts, J. K. (March, 2000). *Nested ANOVA vs. crossed ANOVA: When and how to use which.* Paper presented at the annual meeting of the Educational Research Association, College Station, TX. (ERIC Document Reproduction Service No. ED forthcoming)

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *Sociological Review, 15,* 351-7.

Rock, D. A., Pollack, J. M., & Quinn, P. (1995). *Psychometric report for the NELS:88 base year through second follow-up* (NCES 95-382). Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Seltzer, M. (1991). The use of data augmentation in fitting hierarchical models to educational data. Unpublished doctoral dissertation, University of Chicago.

Seltzer, M. H. (June, 1994). Studying variation in program success: A multilevel modeling approach. *Evaluation Review, 18,* 342-61.

Shah, B. V., Barnwell, B. G., & Bieler, G. S. (1997). *SUDAAN user's manual, release 7.5.* Research Triangle Park, NC: Research Triangle Institute.

Singer, J. D. (1999). Using SAS PROC MIXED to fit multilevel models, hierarchical models, and individual growth models. *The Journal of Educational and Behavioral Statistics, 24,* 323-355.

Snijders, T. A. & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage.

Snijders, T. A. & Bosker, R. J. (1993). Standard errors and sampling sizes for two-level research. *Journal of Educational Statistics, 18*, 237-261.

Thompson, B. (1998, April). *Five methodology errors in educational research: The pantheon of statistical significance and other faux pas*. Invited address presented at the annual meeting of the American Educational Research Association, San Diego. (ERIC Document Reproduction Service No. ED 419 023)

Table 1
Results from single predictor regressions for multilevel and weighted samples.

Variable	Multilevel Sample			OLS Weighted Sample		
	β_1	se	ICC	β_1	se	r^2
Parents attended a school meeting – *BYS37A*	-1.42	0.26	.311	-2.59	0.28	.02
Parents attended a school event - *BYS37D*	-1.79	0.25	.309	-4.09	0.27	.05
Attend science lab at least once a week - *BYS67AA*	-1.06	0.38	.313	-1.70	0.34	.01
Afraid to ask question in science class - *BYS72B*	1.95	0.13	.321	2.46	0.17	.03
Science will be useful in my future - *BYS72C*	-1.44	0.12	.333	-1.33	0.15	.02
Percent minority in school - *G8MINOR*	-1.95	0.12	.119	-1.75	0.06	.15
Percent free lunch in school - *G8LUNCH*	-1.77	0.11	.193	-1.47	0.05	.11
Socio-economic status composite - *BYSES*	4.05	0.17	.182	5.60	0.14	.20
Yearly family income - *BYFAMINC*	0.86	0.05	.234	1.36	0.04	.15
Number of hours spent on homework per week - *BYHOMEWK*	0.93	0.09	.336	1.32	0.10	.04
Number of Hispanic teachers - *BYSC20C*	-1.14	0.22	.318	-0.74	0.09	.01
Number of Black teachers - *BYSC20D*	-1.55	0.14	.246	-1.35	0.05	.10
Number of White teachers - *BYSC20E*	0.92	0.19	.314	0.74	0.07	.02
Science taught in non-English language - *BYSC45B3*	5.68	1.09	.314	4.05	0.38	.02
Belong to a parent-teacher organization - *BYP59A*	-2.19	0.25	.304	-4.35	0.30	.04
Dummy variable - Black students versus rest – *BLACK*	-4.88	0.33	.291	-6.91	0.25	.09
Dummy variable – White students versus rest – *WHITE*	4.45	0.31	.258	7.74	0.25	.15
Dummy variable – Hispanic students versus rest - *HISPANIC*	-2.95	0.35	.318	-3.91	0.28	.02

Note. These are variable names embedded within the NELS:88 dataset.

Table 2
Results comparing the ordinary least squares model and the multilevel model.

Variable	OLS Estimate	Multilevel Estimate	Greater Value
Intercept	46.06 (1.24)	41.91 (1.81)	OLS
Parents attended a school event – *BYS37D*	-2.00 (0.30)	-1.25 (0.27)	OLS
Hours spent on homework per week - *BYHOMEWK*	1.06 (0.11)	0.94 (.083)	OLS
Science taught in non-English language – *BYSC45B3*	0.68 (0.42)	3.07 (.86)	Mult
Belong to a parent-teacher organization – *BYP59A*	-2.07 (0.31)	-1.48 (.27)	OLS
Dummy variable – White students versus rest – *White*	6.58 (0.29)	4.16 (.30)	OLS

Note. These are variable names embedded within the NELS:88 dataset.
Multiple R^2 for the OLS model is 0.194.
Standard error in parenthesis.